9/29/2021

BANK LOAN RISK ANALYSIS-EDA CASE STUDY

PREPARED BY:

- JAYESH JACOB
- SUSANDEEP GANTA

BANK LOAN RISK ANALYSIS- EDA CASE STUDY

TABLE OF CONTENTS:

BANK LOAN RISK ANALYSIS- EDA CASE STUDY

INTRODUCTION:

This case study aims to give an idea of applying EDA in a real business scenario. In this case study, we will develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

PROBLEM STATEMENT:

The loan providing companies find it hard to give loans to the people due to their insufficient or non-existent credit history. Because of that, some consumers use it as their advantage by becoming a defaulter. Suppose you work for a consumer finance company which specializes in lending various types of loans to urban customers. You have to use EDA to analyze the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected. When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile.

Two types of risks are associated with the bank's decision:

1. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.

2. If the applicant is not likely to repay the loan, i.e., he/she is likely to default, then approving the loan may lead to a financial loss for the company.

The data given below contains the information about the loan application at the time of applying for the loan.

READING THE DATA SET:

We have two data sets:

1. Application data set
2. Previous data set

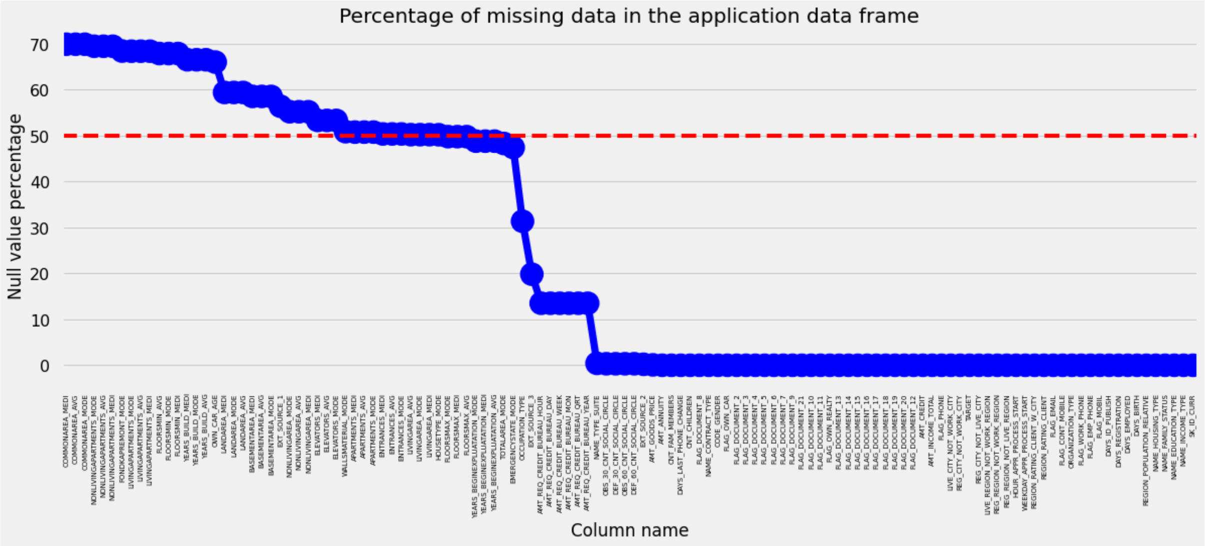
While reading the application data set, we see that the size of the data set is: (307511,122). Which means that there are **307511 rows and 122 columns**.

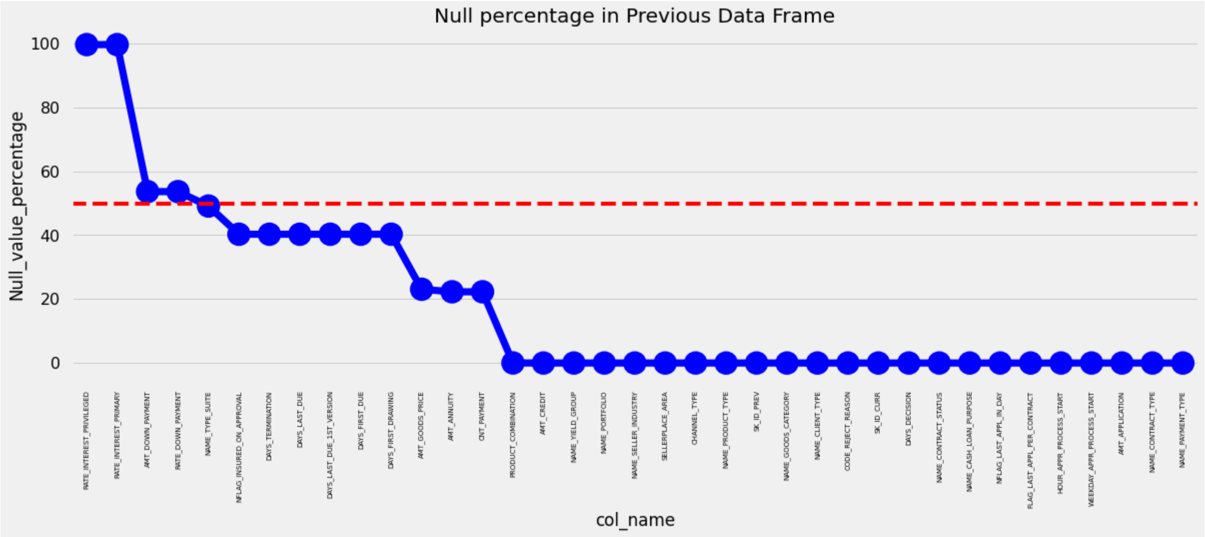
While reading the previous data set, we see that the size of the data set is: (1670214,37). Which means that there are **1670214 rows and 37 columns.**

The sum of **null values** in application data set is 9152466 which is about **24.39%** of the total data that is available.

The sum of null values in previous data set is 11109336 which is about **17.97%** of the total data that is available.

Graph of null value greater than 50%





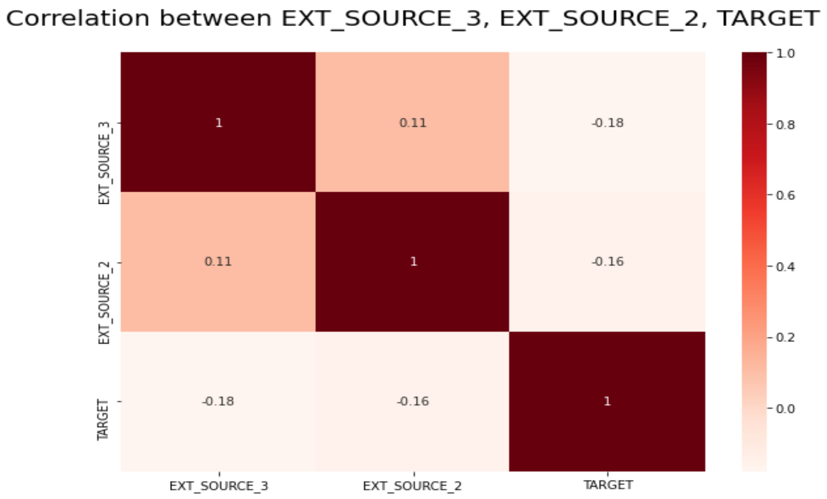
DATA CLEANING AND MANIPULATION:

Since there is almost **25% of null value** in the total data set of the application data frame. Hence, we decided to keep the cutoff range of 50% of null values for each column in the application data set. Therefore, any column which has null value percentage greater than 50% will be dropped from the data frame whereas those columns which had null value lesser than 50% will be imputed with appropriate value.

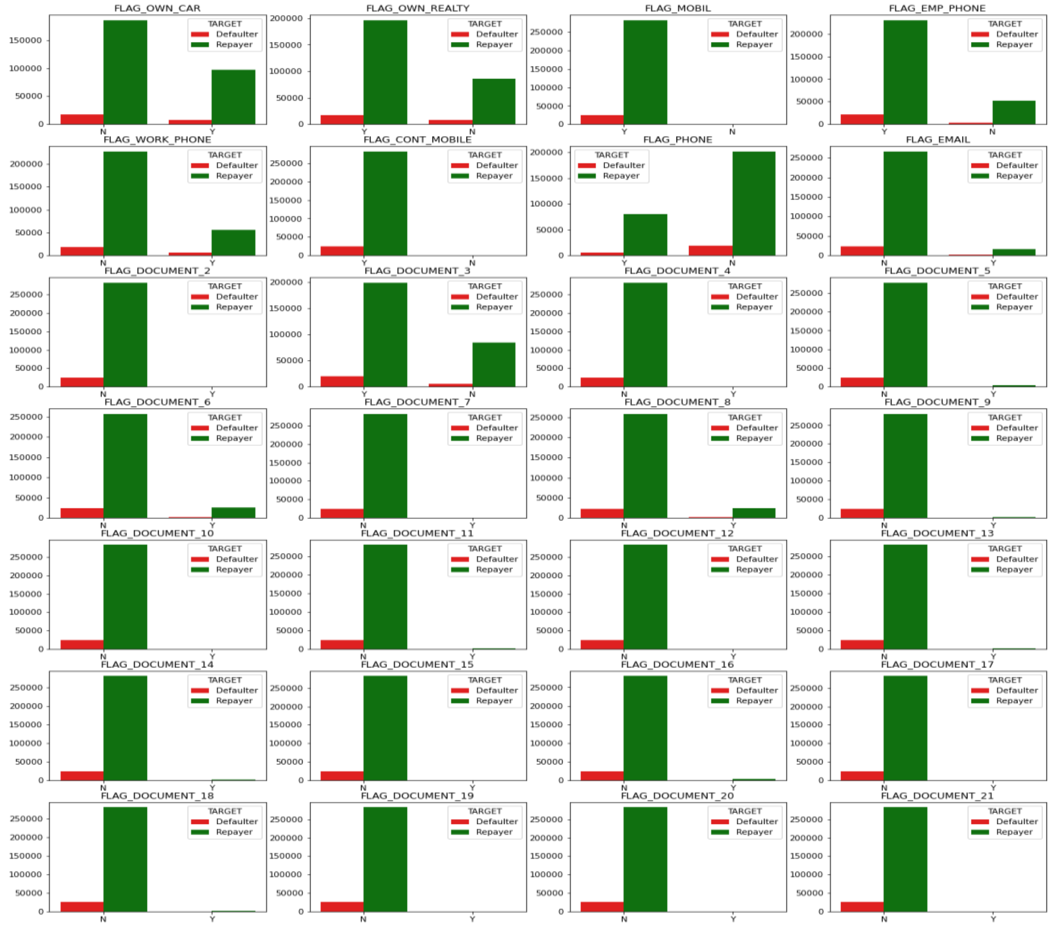
After converting the sum of null values to percentage of null/missing values we see quiet good numbers of columns having null value percentage higher than 50%. Therefore, grouping them all under a column named null_col_50. The **count of null value columns is 41**, similarly the count of columns having null value greater than 50% in previous data frame is **4 columns**. Hence, removing all these columns.

Even after removing the columns with null value greater than 50% we see there are columns having null value around 40% and 30%. Therefore, we plot a heat map to find the correlation between columns and the Target column.

Hence after closely scrutinizing the values and relationship with the TARGET column we concluded that there were few columns which did not establish any prominent relationship. Therefore, decided to drop those columns.



From the above heatmap you can clearly see that there is no evident relation between few columns and Target column. Hence, decided to drop it. The graph below is regarding the various form’s v/s Target column. We can see that apart from FLAG_3 rest all do not show any evident relationship with the target column. Hence, decided to drop these columns as well.



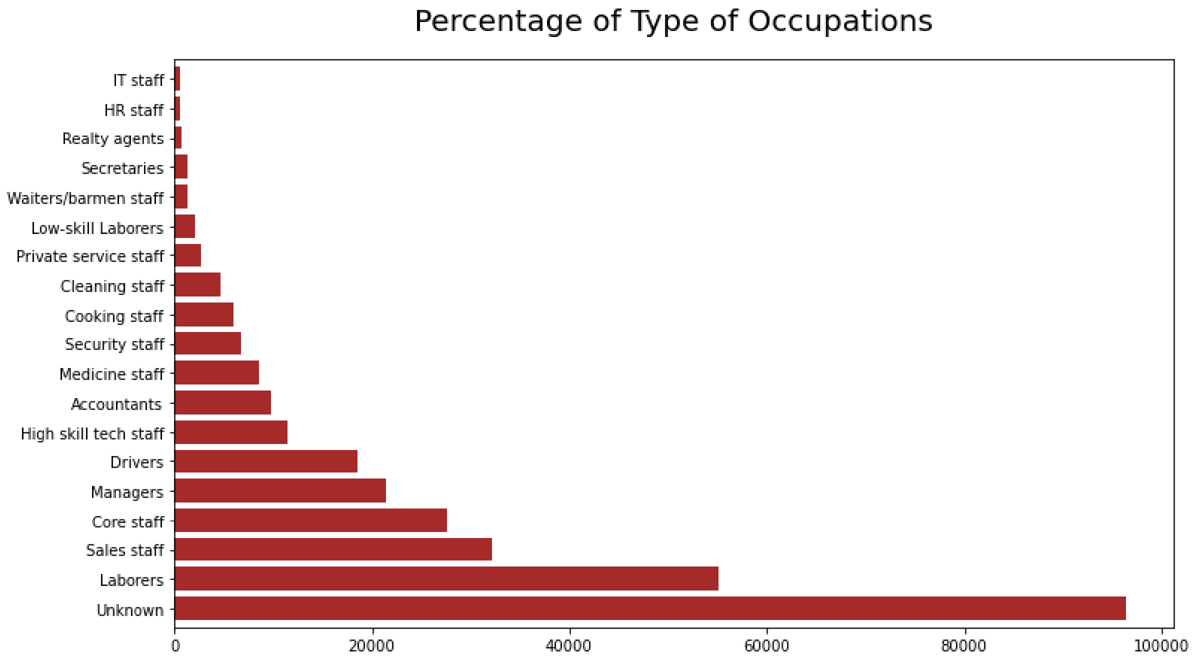
Value Imputation:

After identifying and dropping the columns which prominently does not establish any relation with Target column. We were left with columns where null value percentage was around 20% and less. Hence, imputed all those columns based on their mean, median and mode value.

Rules of Imputation:

1. If the mean and median does not have large margin of difference then impute the missing values with the mean value of the column.
2. If the mean and median does have large margin of difference then impute the missing value with the median value of the column.
3. If the column has "categorical" value then either go with mode or leave the column un-treated.

For example, occupational column had null values which was replaced by 'unknown' value.



Standardizing the values:

1. Columns like AMT_INCOME_TOTAL, AMT_CREDIT, AMT_GOODS_PRICE have higher values. It will be convenient if we convert them to categorical form which would decide the range.
2. Columns like DAYS_BIRTH, DAYS_EMPLOYED, DAYS_ID_PUBLISH, DAYS_REGISTRATION, DAYS_LAST_PHONE_CAHNGE has negative values. Days cannot be in negative form therefore converting them all to positive value.

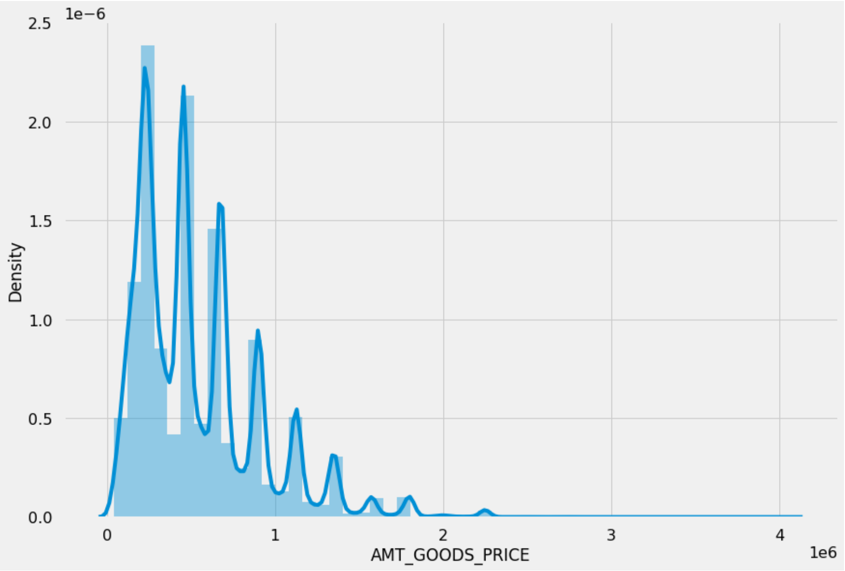
3. Convert DAYS_BIRTH to AGE in years, DAYS_EMPLOYED to YEARS EMPLOYED.

Steps taken to Impute values:

1. Creating bins for Income amount as AMT_INCOME_RANGE column.

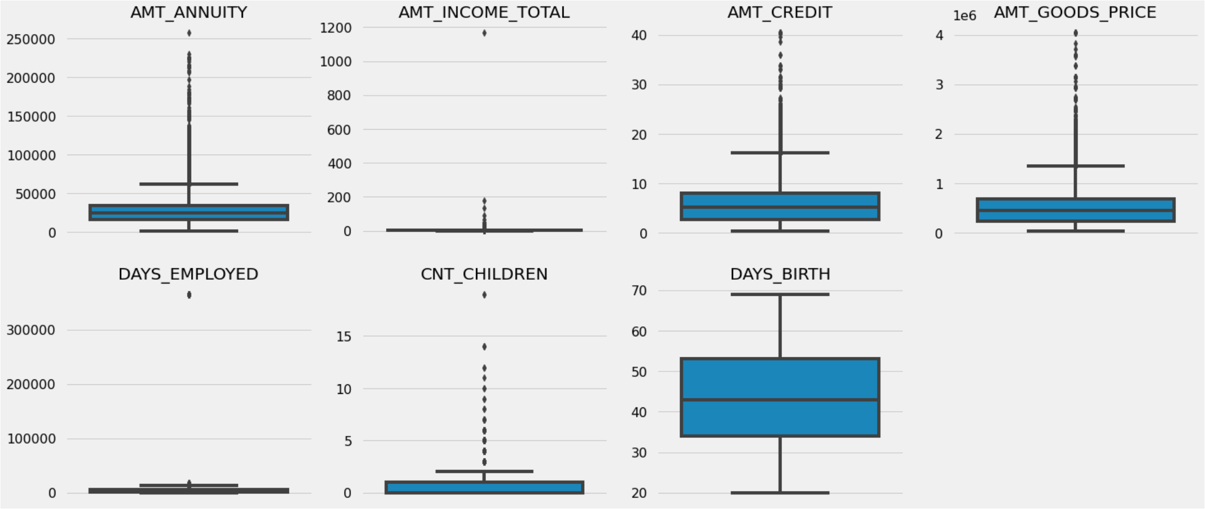
2. Creating bins for AMT_CREDIT as AMT_CREDIT_RANGE column.

3. Creating bins for AMT_GOODS_PRICE as AMT_CREDIT_RAGE column.



Finding the Outliers:

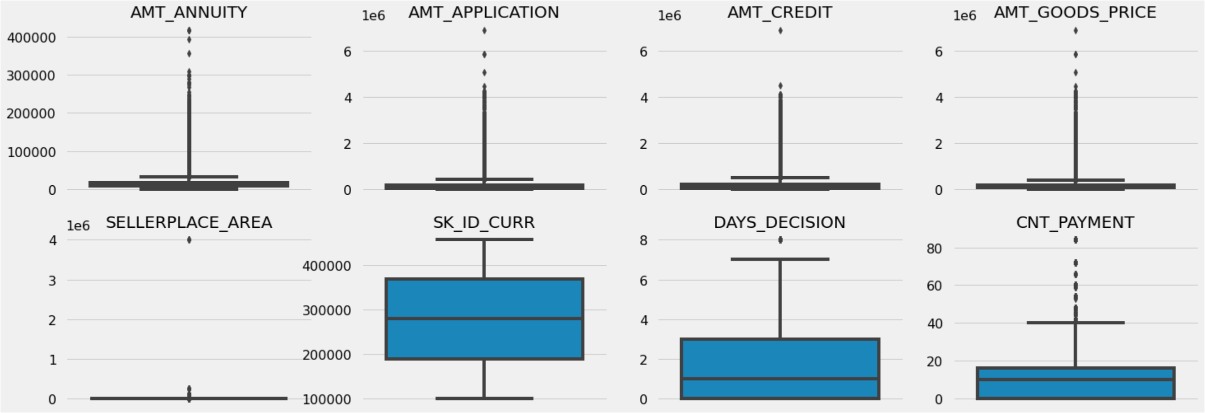
From the **application data frame**, we have selected few columns which has extreme values for max. Hence decided to plot boxplot to find the outliers in those columns.



Inferences of boxplot:

1. Days_birth column does not have any outlier; therefore, we can totally rely upon these values for the analysis.
2. Days_employed and amt_income_total has huge outliers, which indicated that days_employed has wrong values and amt_income_total has few applicants whose salary is higher than the mean salary of the column.
3. Cnt_children have outliers, which makes it very suspicious that some of the applicants have given children count more than 10.

From the **previous data frame**, we have selected few columns which has extreme values for max. Hence decided to plot boxplot to find the outliers in those columns.



Insight: It can be seen that in previous application data

1. AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA have huge number of outliers.

2. CNT_PAYMENT has few outlier values.

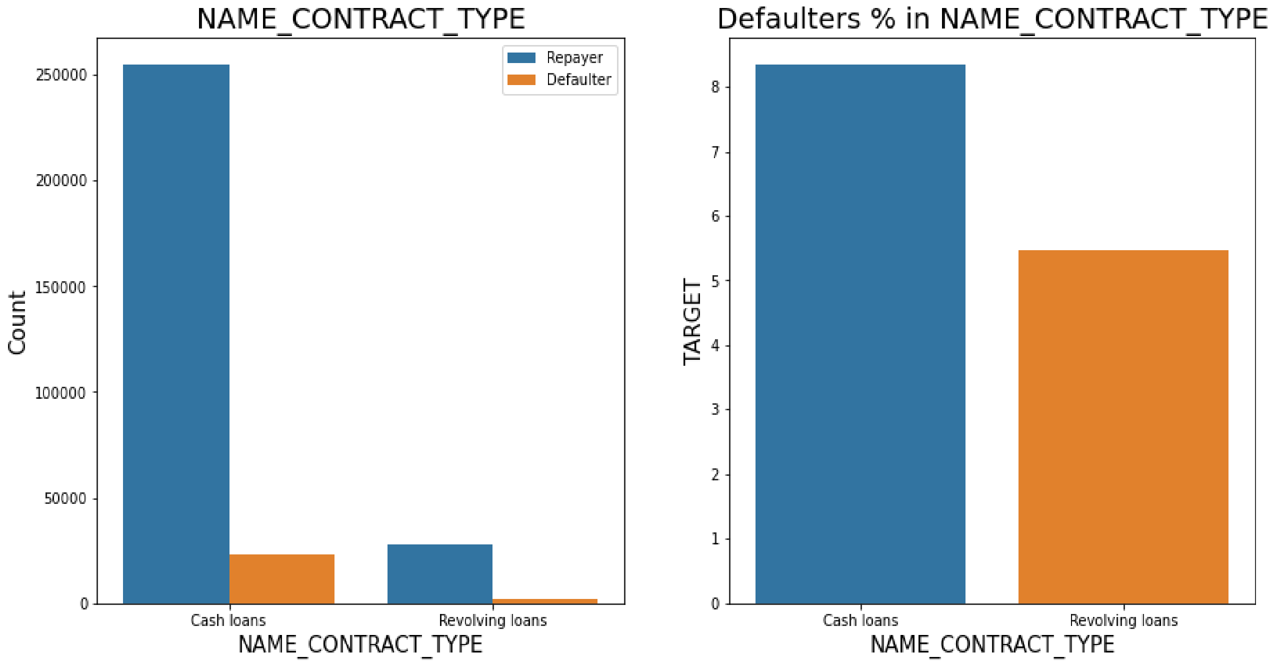
3. SK_ID_CURR is an ID column and hence no outliers.

4. DAYS_DECISION has little number of outliers indicating that these previous applications decisions were taken long back.

DATA ANALYSIS:

Segmented univariate analysis:

1. **NAME_CONTRACT_TYPE**: **This is the column where we see have 2 values, which is cash loan and revolving loan.**

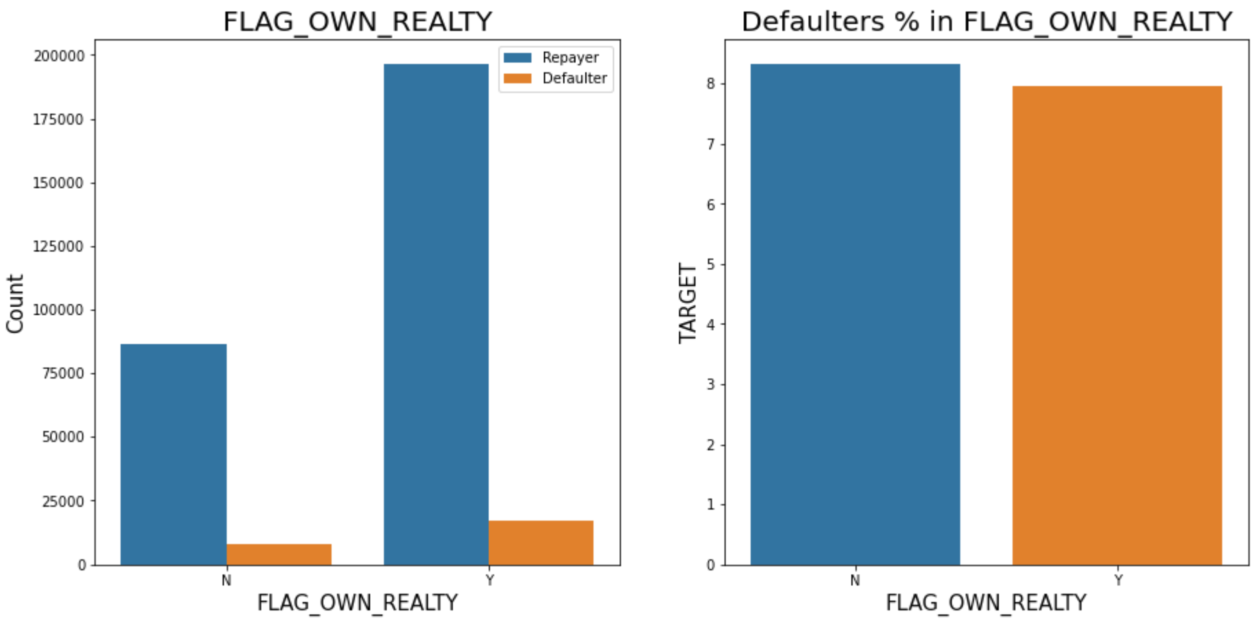


Inferences: For Contract type:

1.Revolving loans are just a small fraction (10%) from the total number of loans.

2.Around 8-9% Cash loan applicants and 5-6% Revolving loan applicant are in defaulters.

2. **FLAG_OWN_REALTY: This is variable where we are trying to find the relationship between people owning property v/s target which has parameters of defaulters and re-payers.**

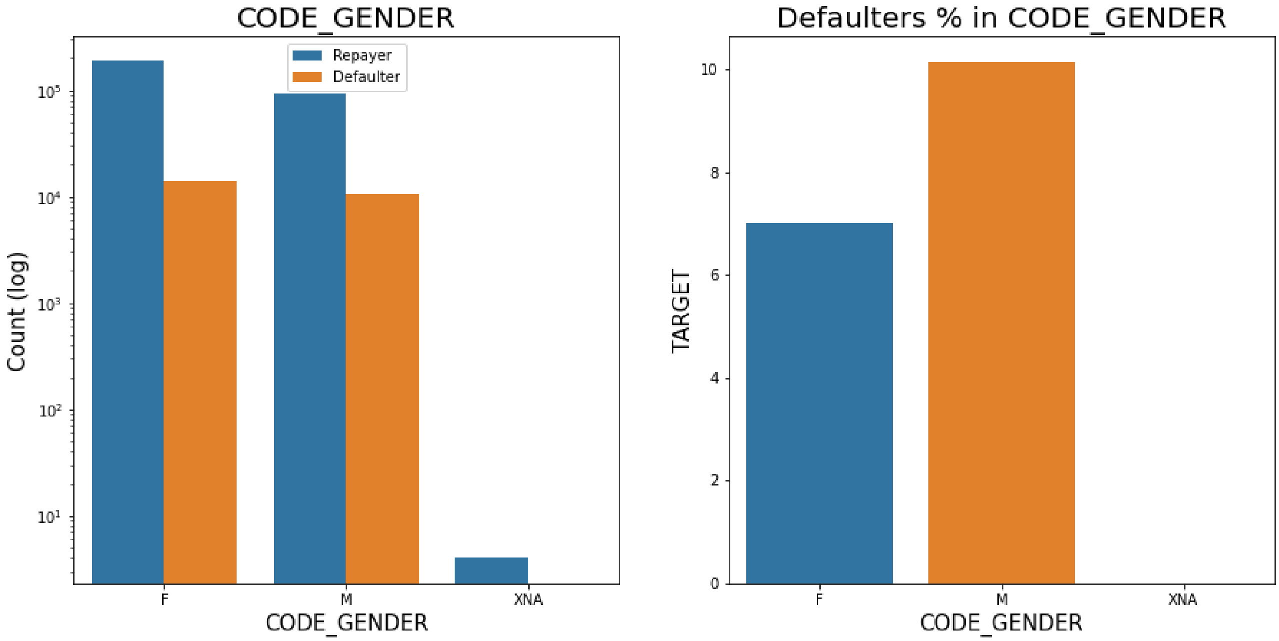


Inferences: For the case of owning Real Estate:

1. The clients who own real estate are more than double of the ones that don't own.
2. The defaulting rate of both categories are around the same (~8%). Thus, we can infer that there is no correlation between owning a reality and defaulting the loan.

BANK LOAN RISK ANALYSIS- EDA CASE STUDY

3. **Checking the type of Gender on loan repayment status. Plotting a graph of "CODE_GENDER" v/s "TARGET". Trying to analyse which gender tends to default the loan payment.**

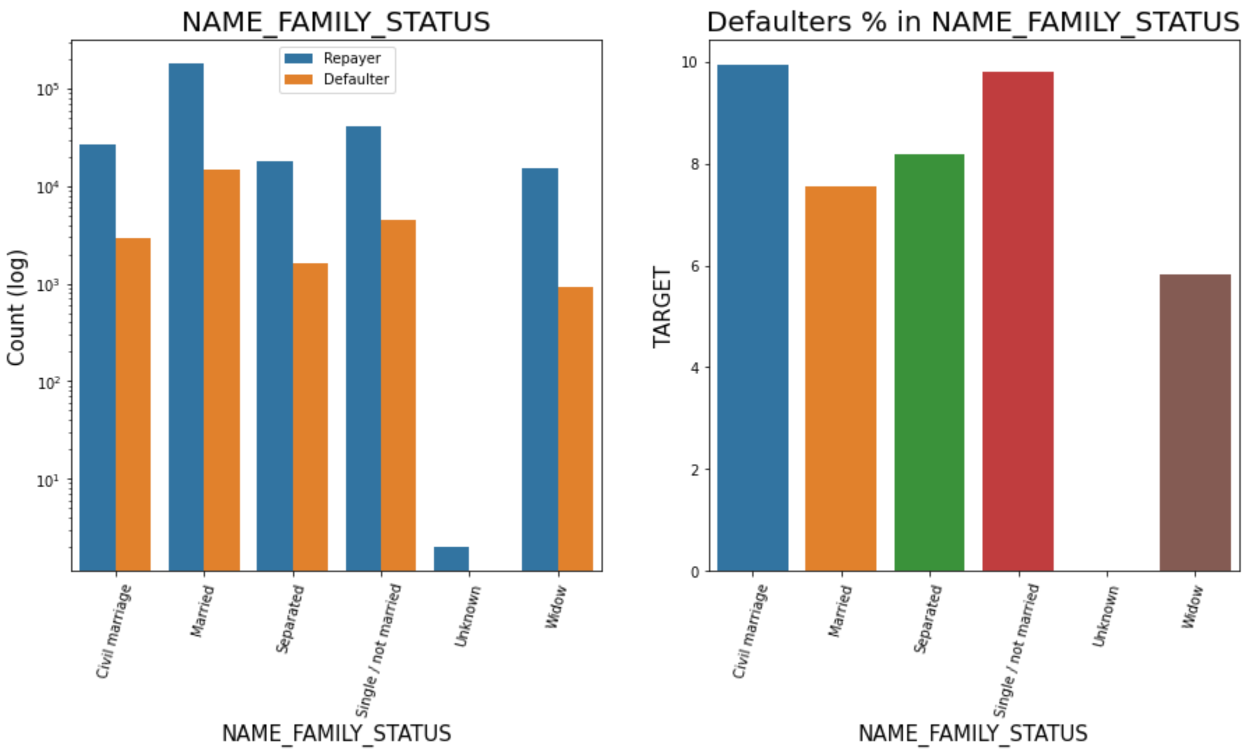


Inferences: For Gender Type:

1. 1.The number of female clients is almost double the number of male clients.
2. Based on the percentage of defaulted credits, males have a higher chance of not returning their loans.
3. XNA are the group where they take the loan and repay it. There is no sign of defaulter from their gender group.

4. **Analysing Family status based on loan repayment status. Since this is a categorical column and we are trying to find which family status tend to take more loan, which family status tend to default the loan and which family status has higher percentage of loan repayment.**

Plotting a graph of Family status v/s Target

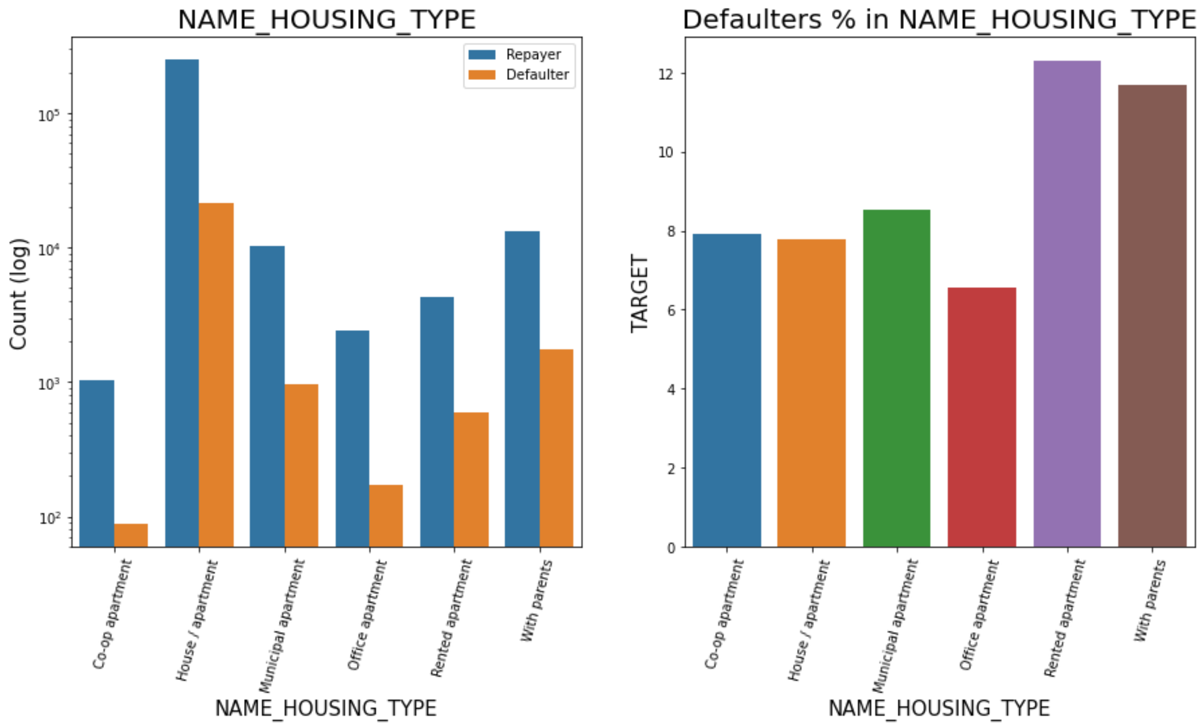


Inferences: For Family Status:

1. Most married people have taken loan, followed by Single/Not married.
2. In Percentage of defaulters, Civil marriage has the highest percentage of defaulting.
3. Large number of Single/not married take loan and then they are the 2nd largest defaulter.

5. **Analysing Housing Type based on loan repayment status. Here we are trying to find out which segment of housing type are defaulters and which segment repays the loan on time.**

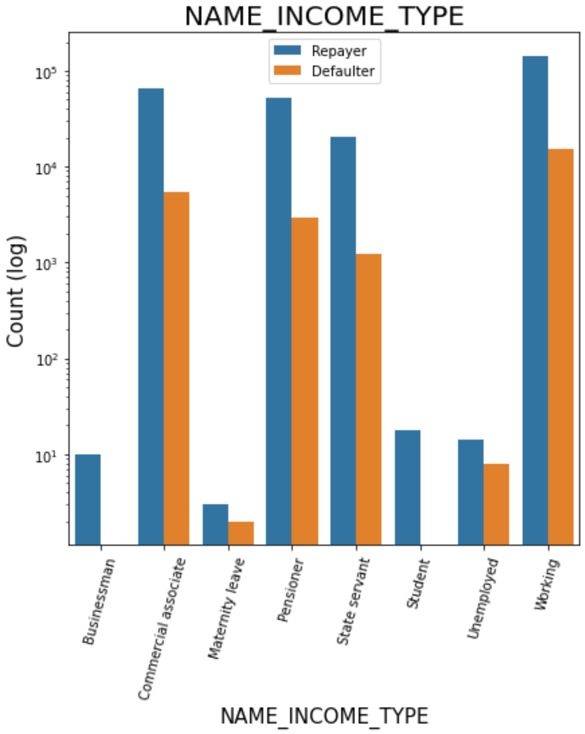
   Plotting graph of NAME_HOUSING_TYPE v/s TARGET

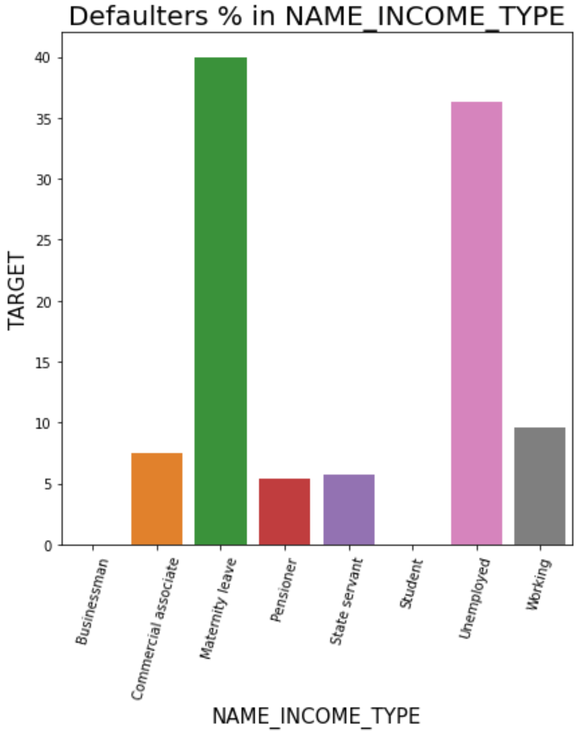


Inferences: For Applicant House type:

1. Majority of people live in House/apartment.
2. People living in office apartments have lowest default rate.
3. People living with parents and living in rented apartments have higher probability of defaulting.
4. People living in House/apartment are good percentage of re-payer.
5. People living in Co-op apartment have less ratio of default.

6. **Analysing Income Type based on loan repayment status. By this we are trying to find which income type repay on time on which group of income type is risky for the business.**

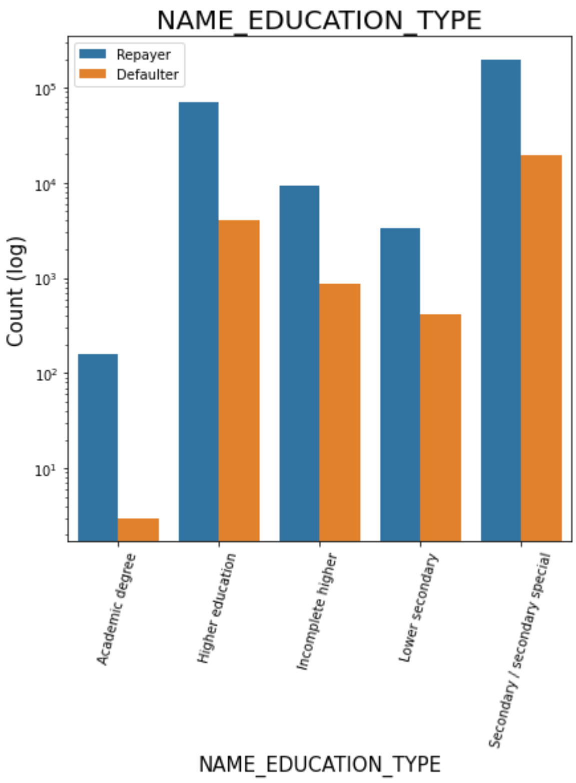


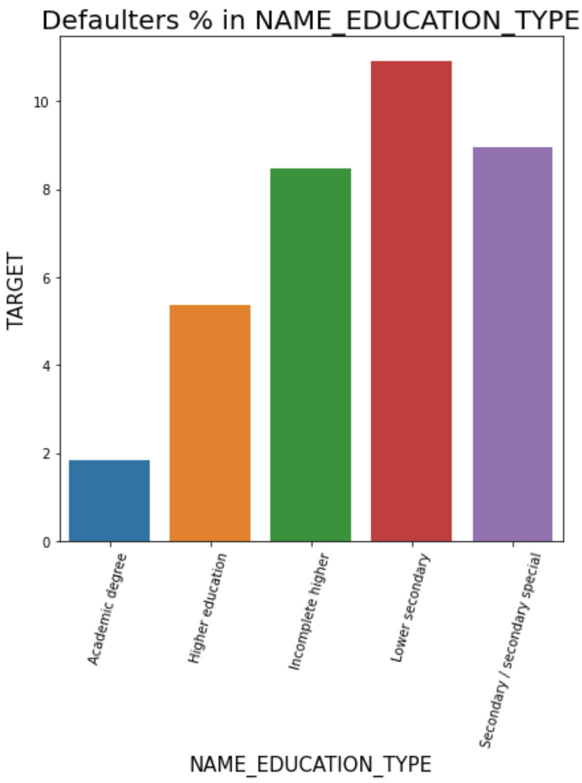


Inferences: For Income Type:

1. Most of applicants for loans income type is Working, followed by Commercial associate, Pensioner and State servant.
2. The applicants who are on Maternity leave have a high defaulting rate, followed by Unemployed.
3. Student and Businessmen though less in numbers, do not have default record. Safest two categories for providing loan.
4. The defaulter ratio of maternity leave and unemployed is high. Hence providing loan to this category is really risky.

7. **Analysing Education Type based on loan repayment status. This category helps us to judge about which group to approach for giving out the loan and expect duly repayment of the loan.**

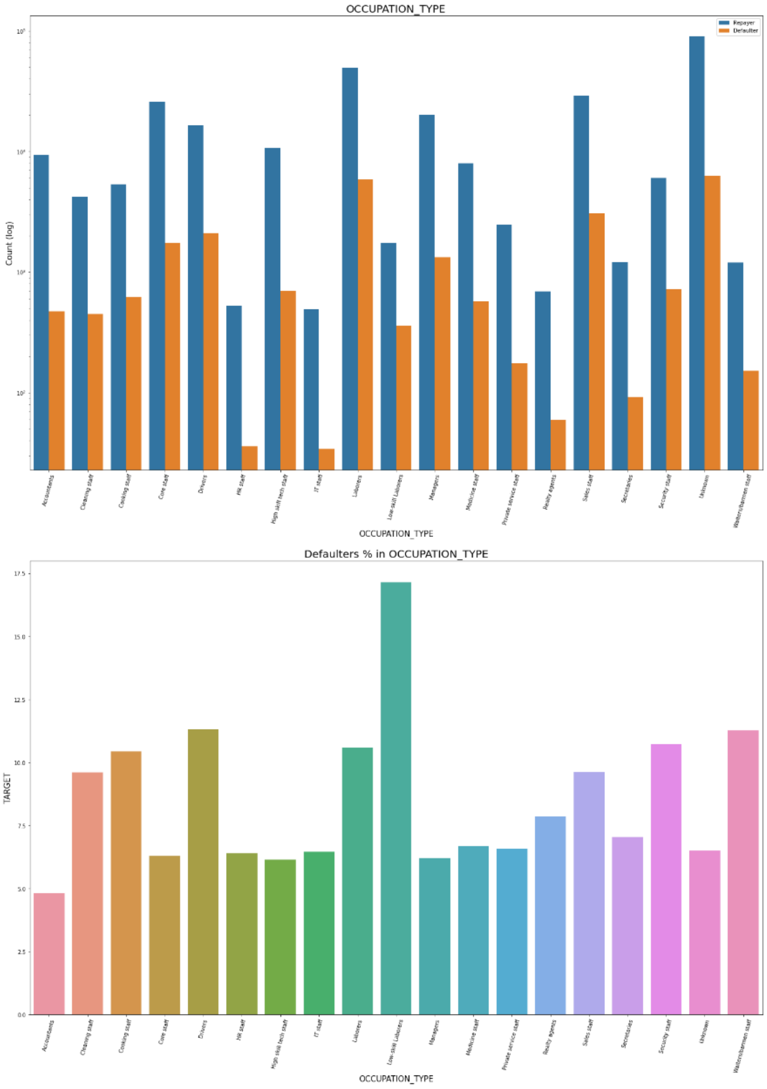




Inferences: For Education Type:

1. Majority of clients have Secondary/secondary special education, followed by clients with Higher education.
2. Very few clients have an academic degree
3. Lower secondary category has highest rate of defaulting around 11%.
4. People with Academic degree are least likely to default.

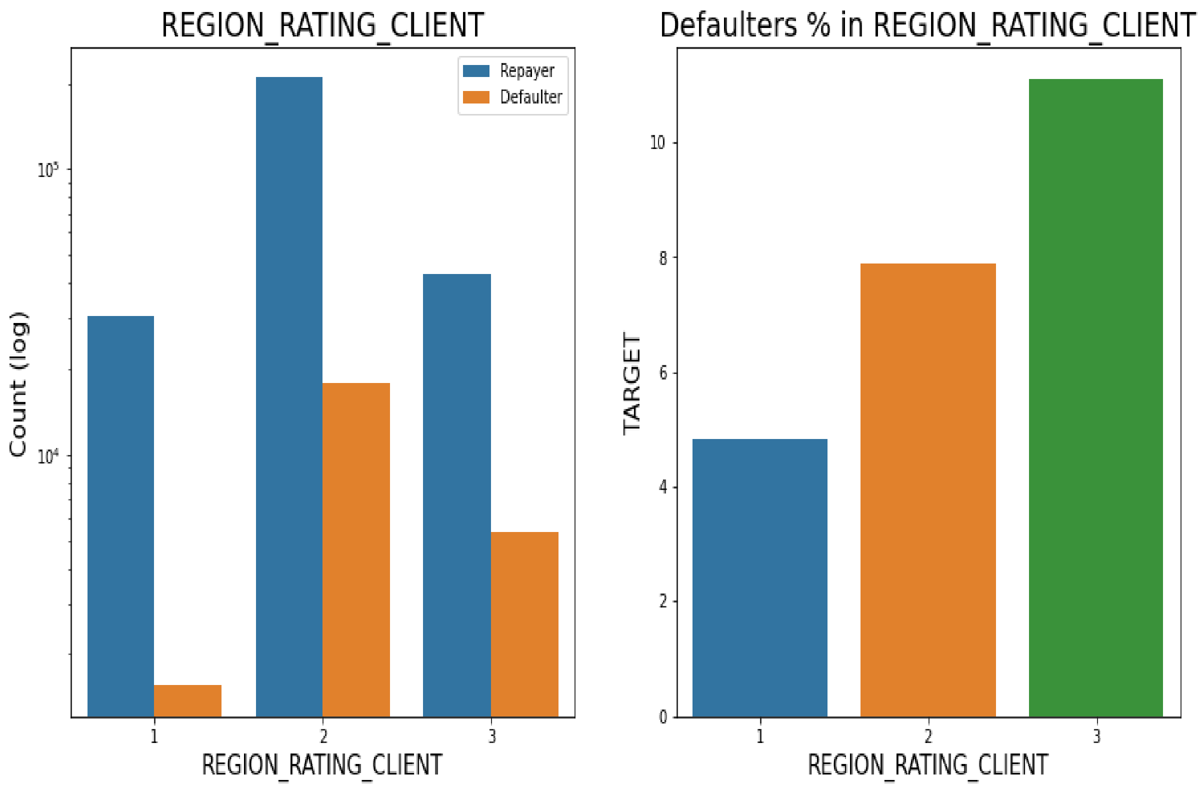
8. Analysing Occupation Type where applicant lives based on loan repayment status.



Inferences: For Occupation Type:

1. Most of the loans are taken by Laborers, followed by Sales staff.
2. IT staff are less likely to apply for Loan.
3. Category with highest percent of defaulters are Low-skill Laborers, followed by Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff.

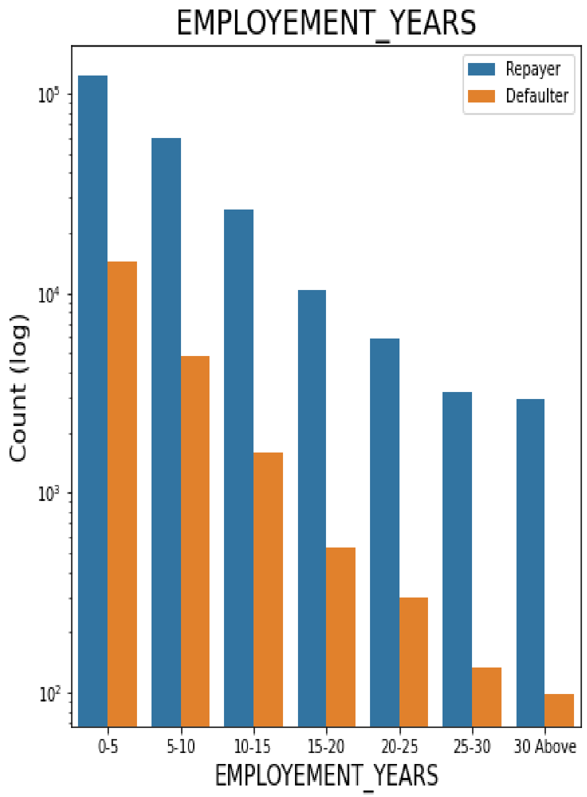
9. **Analysing Region rating where applicant lives based on loan repayment status.**

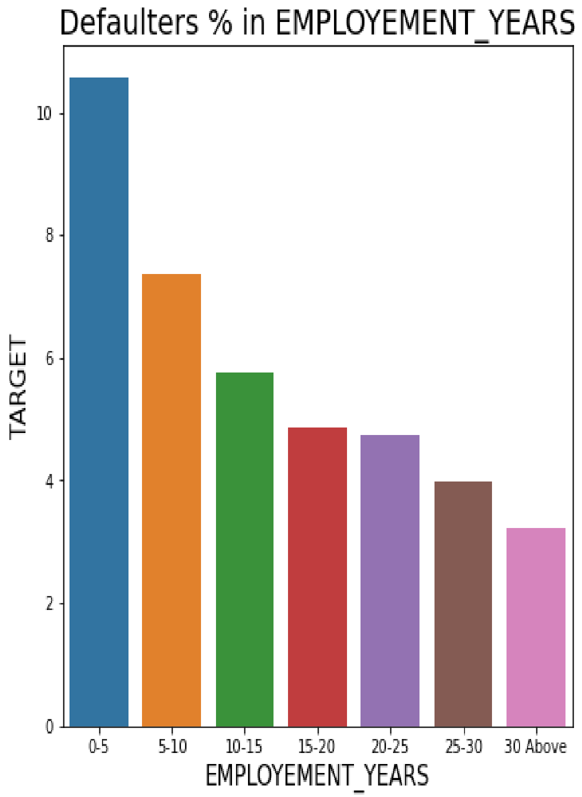


Inferences: For Client Region Rating:

1. Most of the applicants are living in Region with Region_Rating 2.
2. Region Rating 3 has the highest default rate.
3. Applicant living in Region_Rating 1 has the lowest probability of defaulting, thus safer for approving loans.

10. Analysing Employment Year based on loan repayment status. This would help us to understand that whether we should give loan to a person who has just started working on the other hand we should also see there is a group of applicants who are about to hit the age of superannuation and is it profitable to give loan to them.

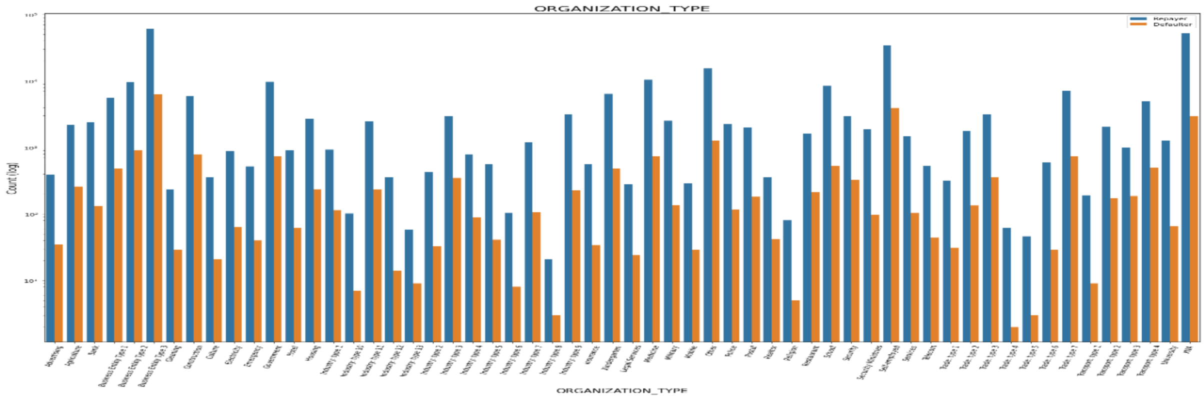


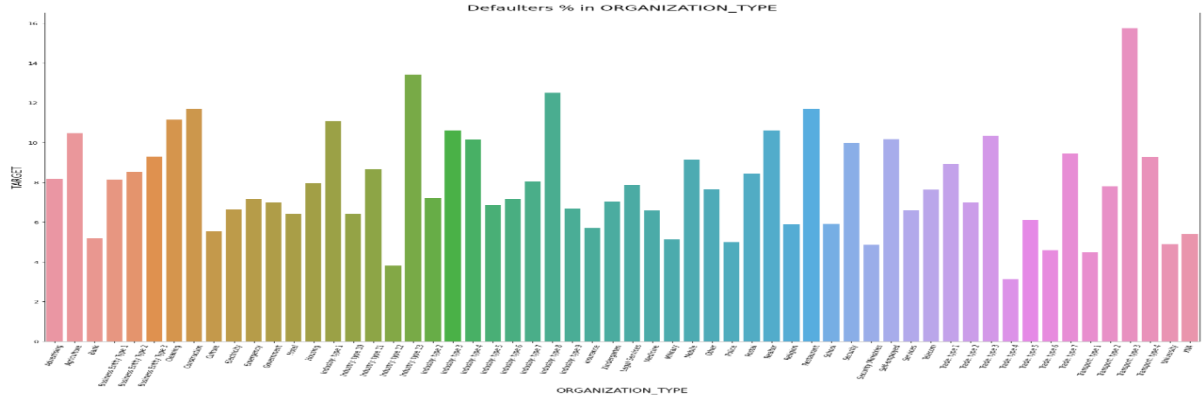


Inferences: For Employment in Years:

1. Majority of the applicants having working experience between 0-5 years are defaulters.
2. With increase of employment year, defaulting rate is gradually decreasing.
3. The default ratio is very low for the applicant who are going to attain the age of superannuation. Its safe to give loan to them.

11. **Checking Loan repayment status based on Organization type. As we previously saw that it is safe to give loans to business person but it is also mandatory to include the type of business, industry etc.**





Inferences: For Organization Type:

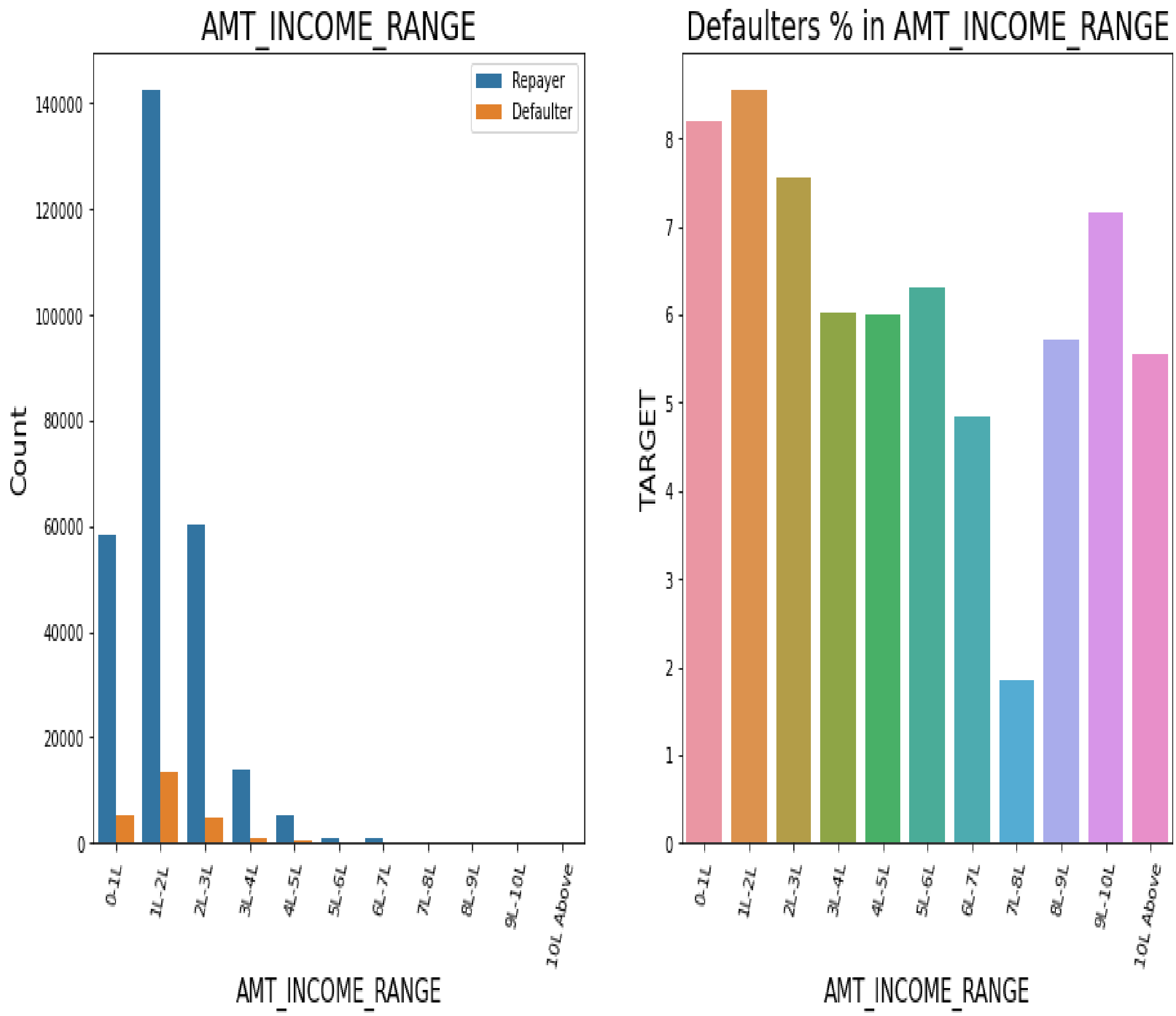
1. Organizations with highest percent of defaultess is Transport: type 3.
2. Self employed people have relative high defaulting rate,to be safer side loan disbursement should be avoided or

12. provide loan with higher interest rate to mitigate the risk of defaulting.

3. For a very high number of applications, Organization type information is unavailable(XNA)
4. It can be seen that following category of organization type has lesser defaulters thus safer for providing loans:

Trade Type 4 and 5, Industry type 8.

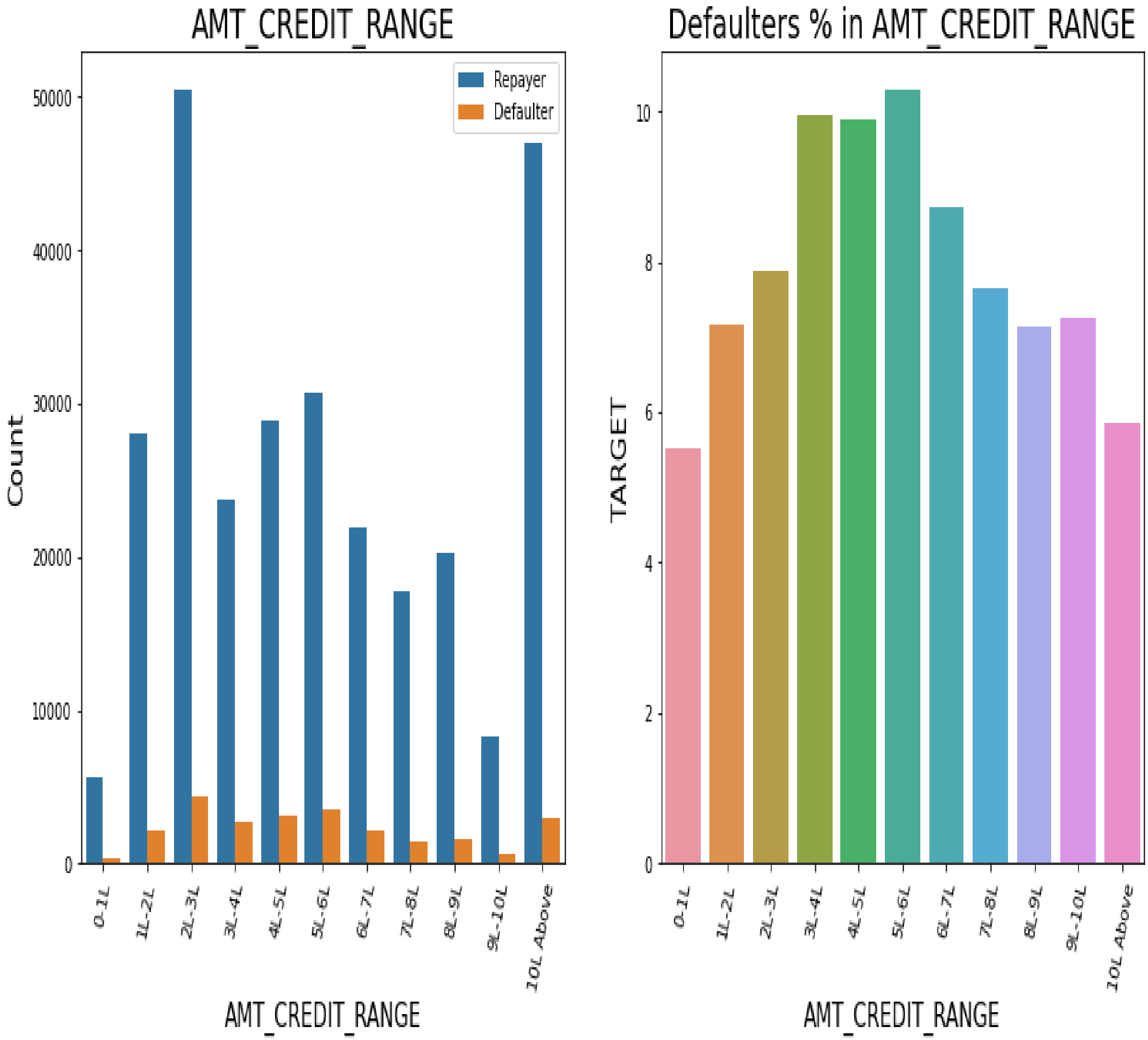
12. Analyzing Amount_Income Range based on loan repayment status.



Inferences: For Applicant Income:

1. Majority of the applications have Income total less than 3 Lakhs.
2. Application with Income less than 3 Lakhs has high probability of defaulting
3. Applicant with Income 7-8 Lakhs are less likely to default.

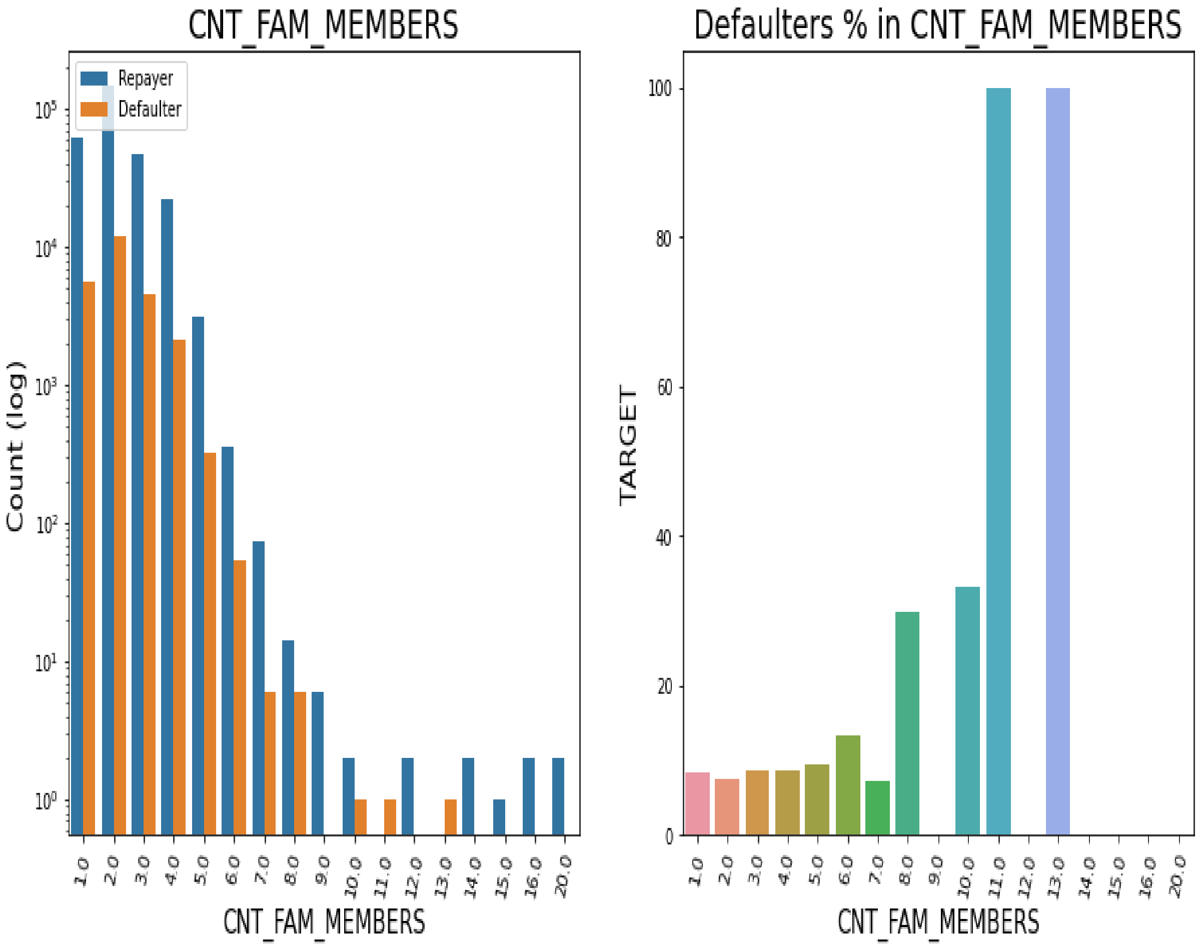
13. Analysing Amount_Credit based on loan repayment status.



Inferences: For Loan Amount:

1. There are high number of applicants have loan in range of 2-3 Lakhs followed by 10 Lakh above range.
2. People who get loan for 3-6 Lakhs have the greatest number of defaulters than other loan range.

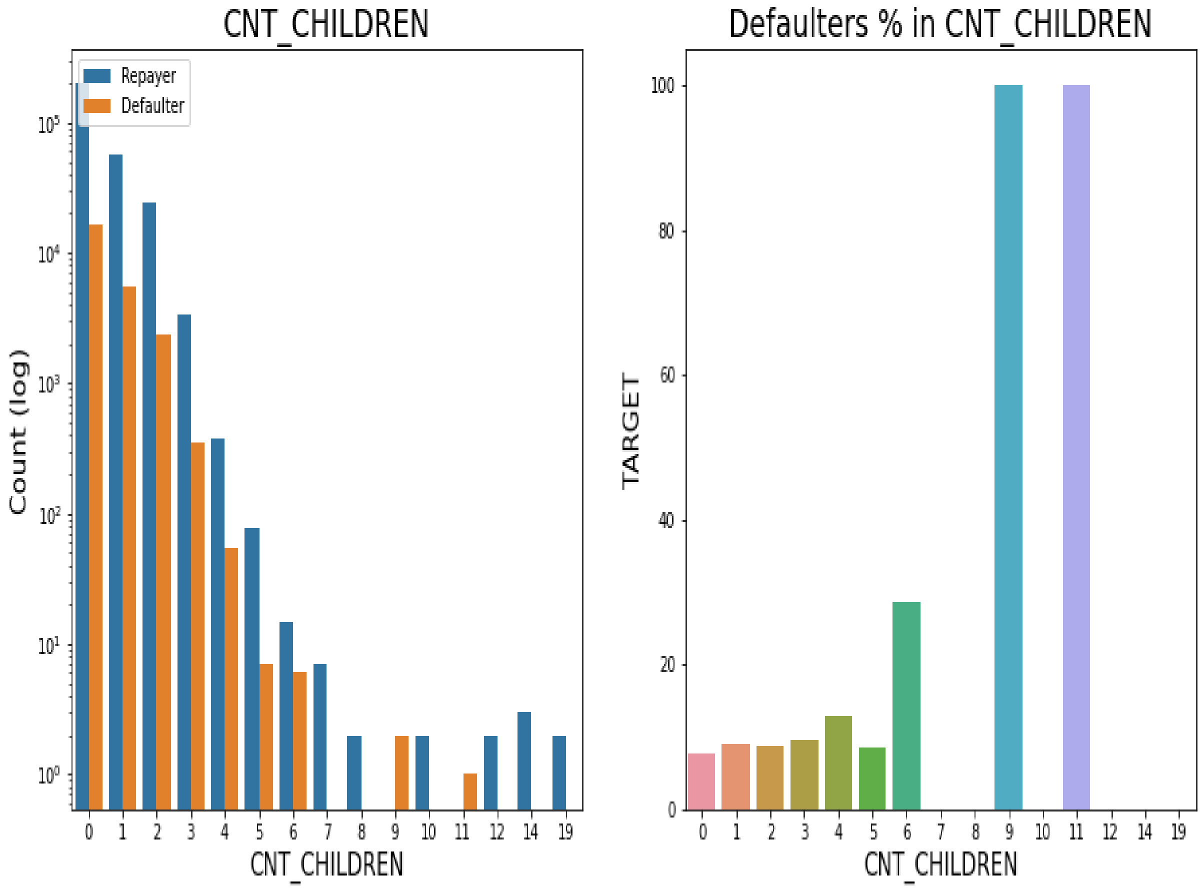
14. Analysing Number of family members based on loan repayment status.



Inferences: For Family Members Count:

1. Families with lesser members are at lower risk of defaulting.
2. Families with 11 or 13 members are the highest risk of defaulting.
3. But for a value of 12 in between these values the default rate is 0.
4. Family members with 14 and above have no sign of defaulters, which is a safe group to provide the loan.

15. Analysing Number of children based on loan repayment status.

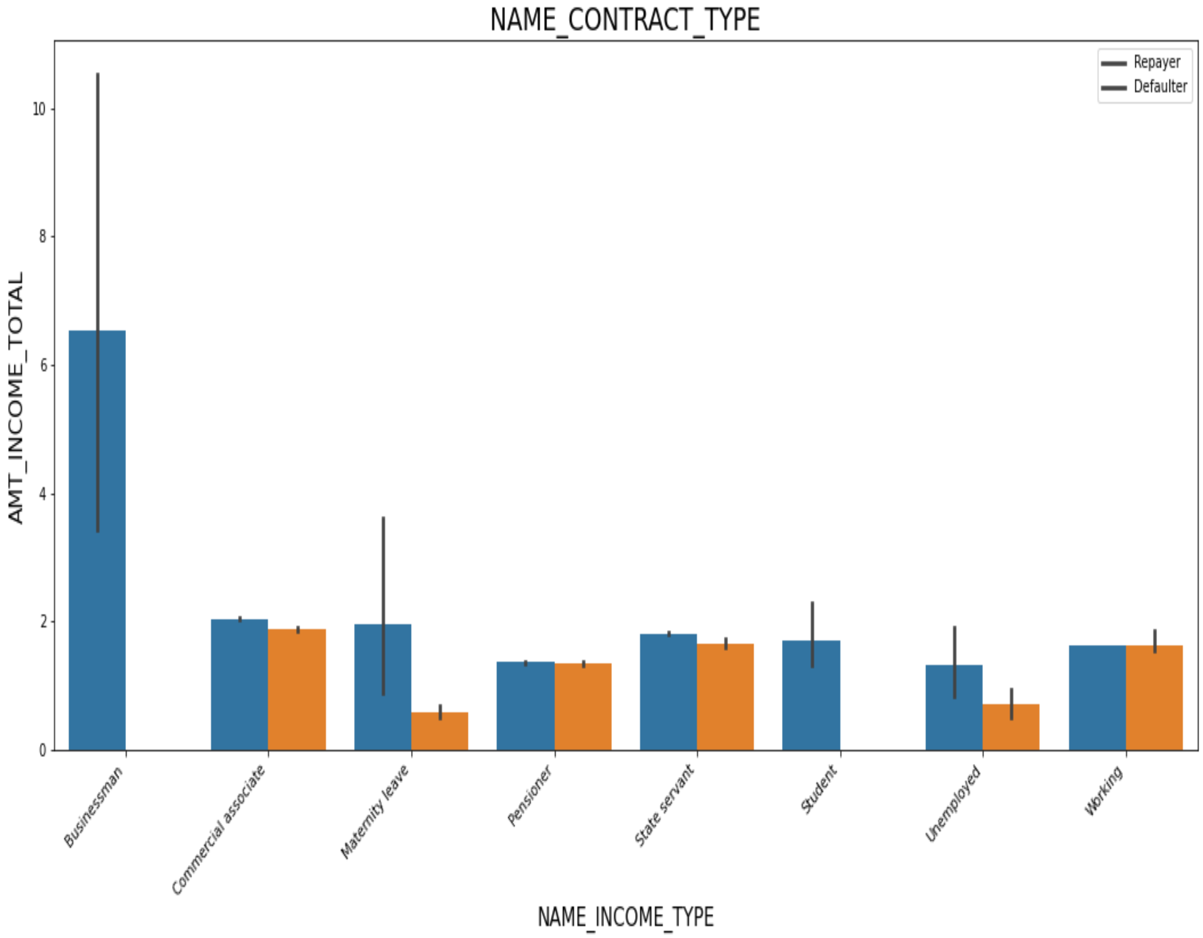


Inferences: Client Children's Count:

1. Most of the applicants do not have children.
2. Clients with 9 or 111 children have a 100 % default rate.
3. Clients with 7 or 8 children have the least default rate.
4. Clients above 12 children’s have 0% default rate.
5. Most applicant have 0 or 1 child and they are the least defaulter compared to all other defaulters.

Categorical Bivariate Analysis:

Income type vs Income Amount Range.

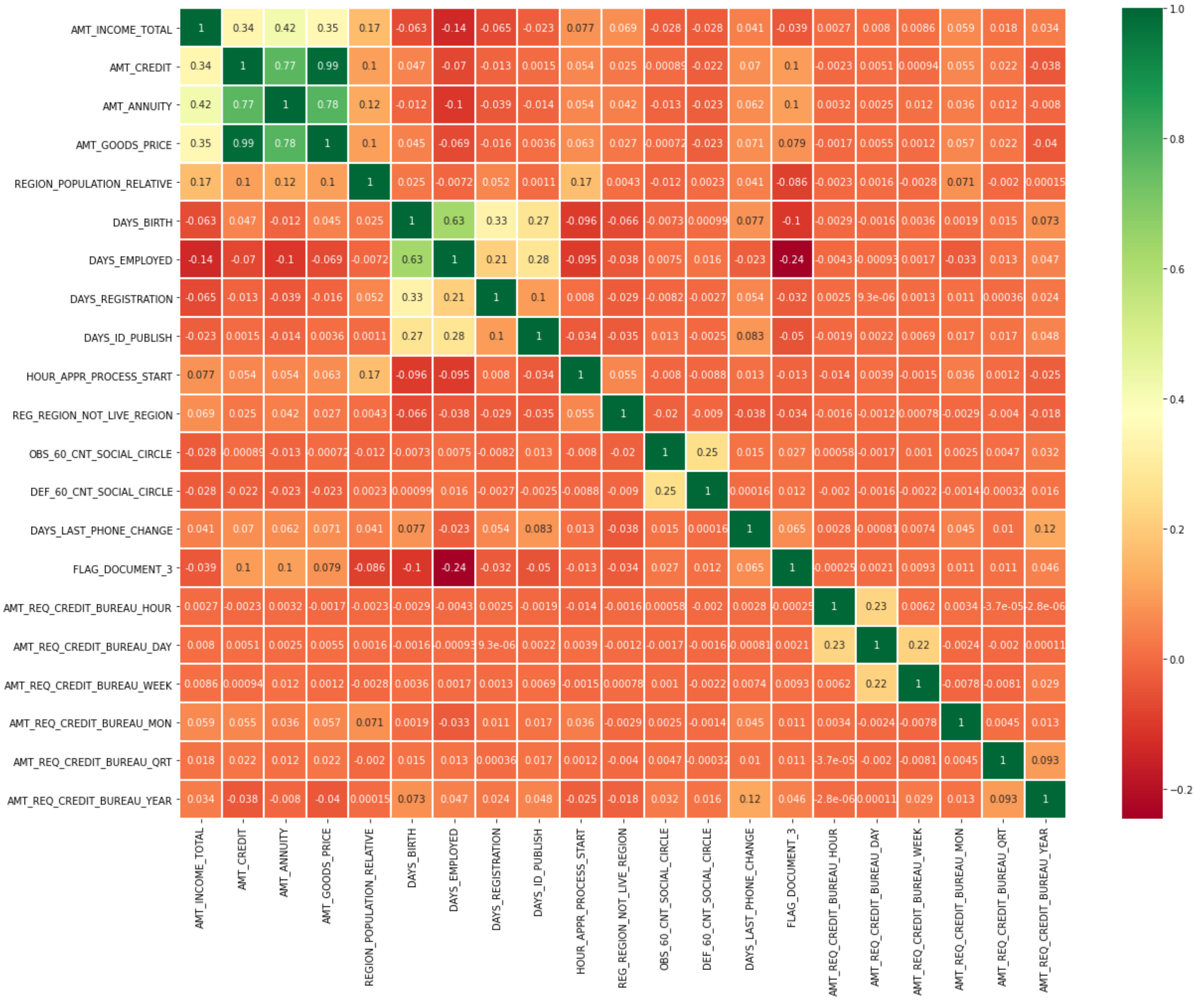


Inferences:

It can be seen that Businessman income is the highest and the estimated range with default 95% confidence level seem to indicate that the income of a Businessman could be in the range of slightly close to 4 lakhs and slightly above 10 lakhs.

Numeric Variables Analysis:

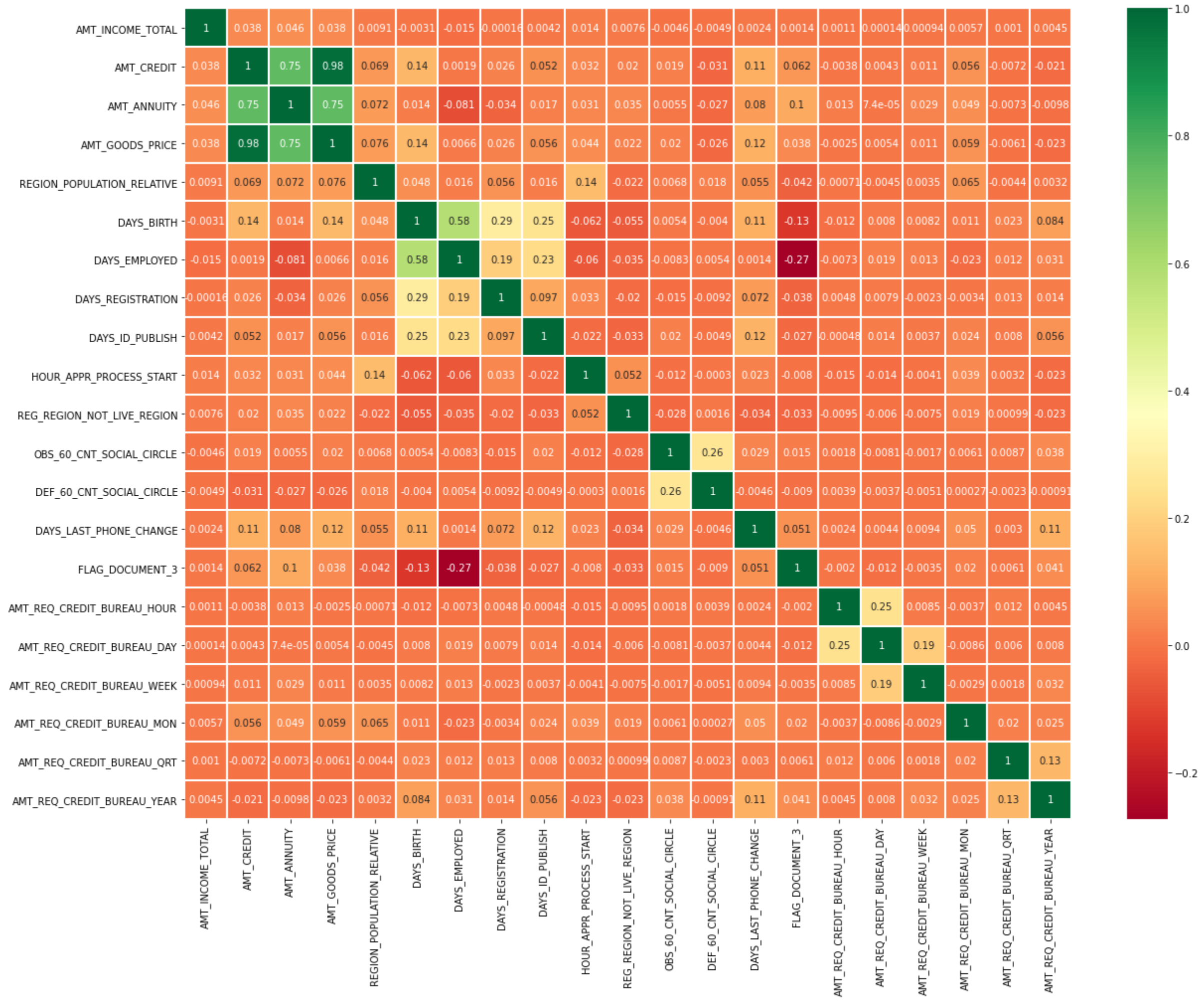
Plotting heatmap to see linear correlation among Re-payers.



Inferences: Correlating factors amongst re-payers:

1. Credit amount is highly correlated with: Goods Price Amount, Loan Annuity, Total Income.
2. We can also see that re-payers have high correlation in number of days employed.

Plotting heatmap to see linear correlation among Defaulters.

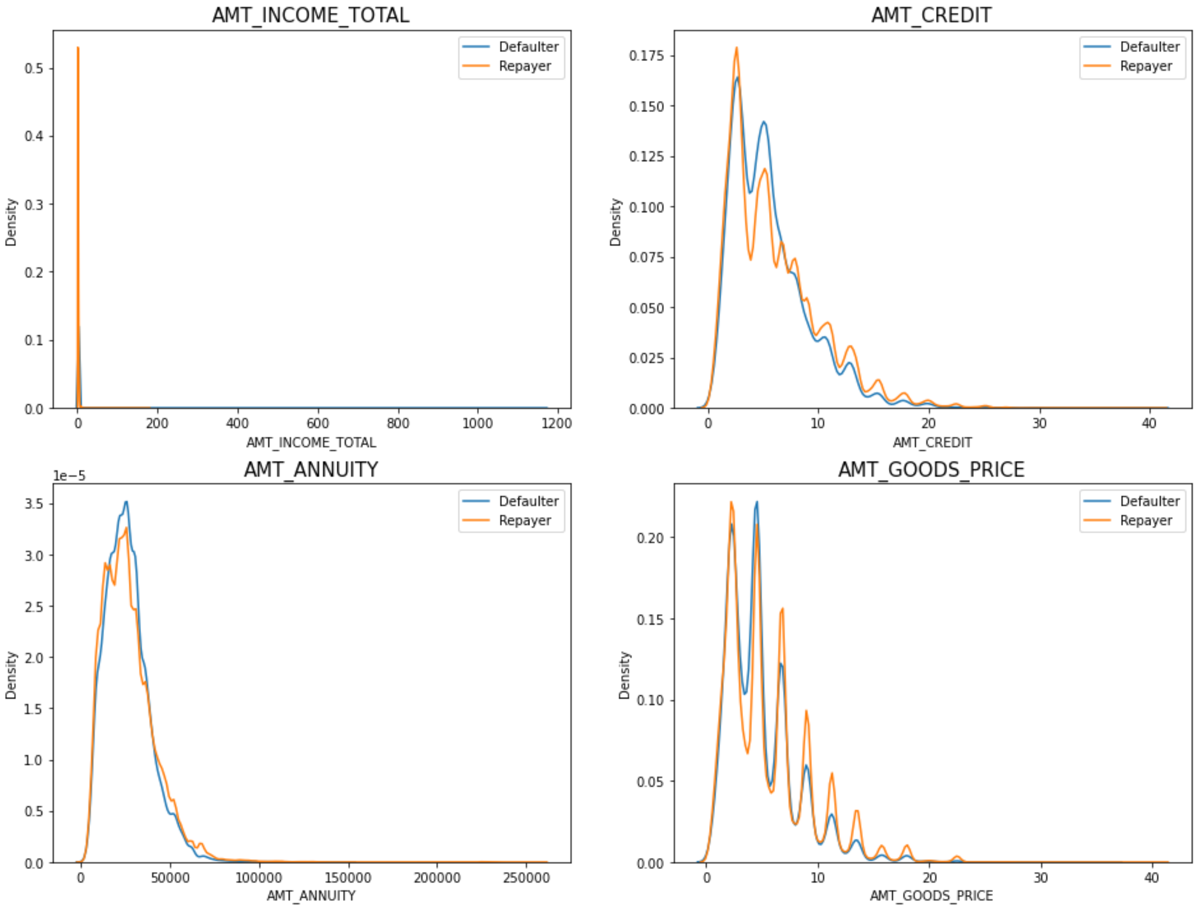


Inferences: Correlating factors among Defaulters:

1. 1.Credit amount is highly correlated with good price amount which is same as re-payers.
2. Loan annuity correlation with credit amount has slightly reduced in defaulters when compared to re-payers
3. We can also see that re-payers have high correlation in number of days employed when compared to defaulters.
4. There is a severe drop in the correlation between total income of the client and the credit amount (0.038) amongst defaulters whereas it is 0.342 among re-payers.

1.1 **Numerical Univariate Analysis:**

Plotting the numerical columns related to amount as distribution plot to see density.

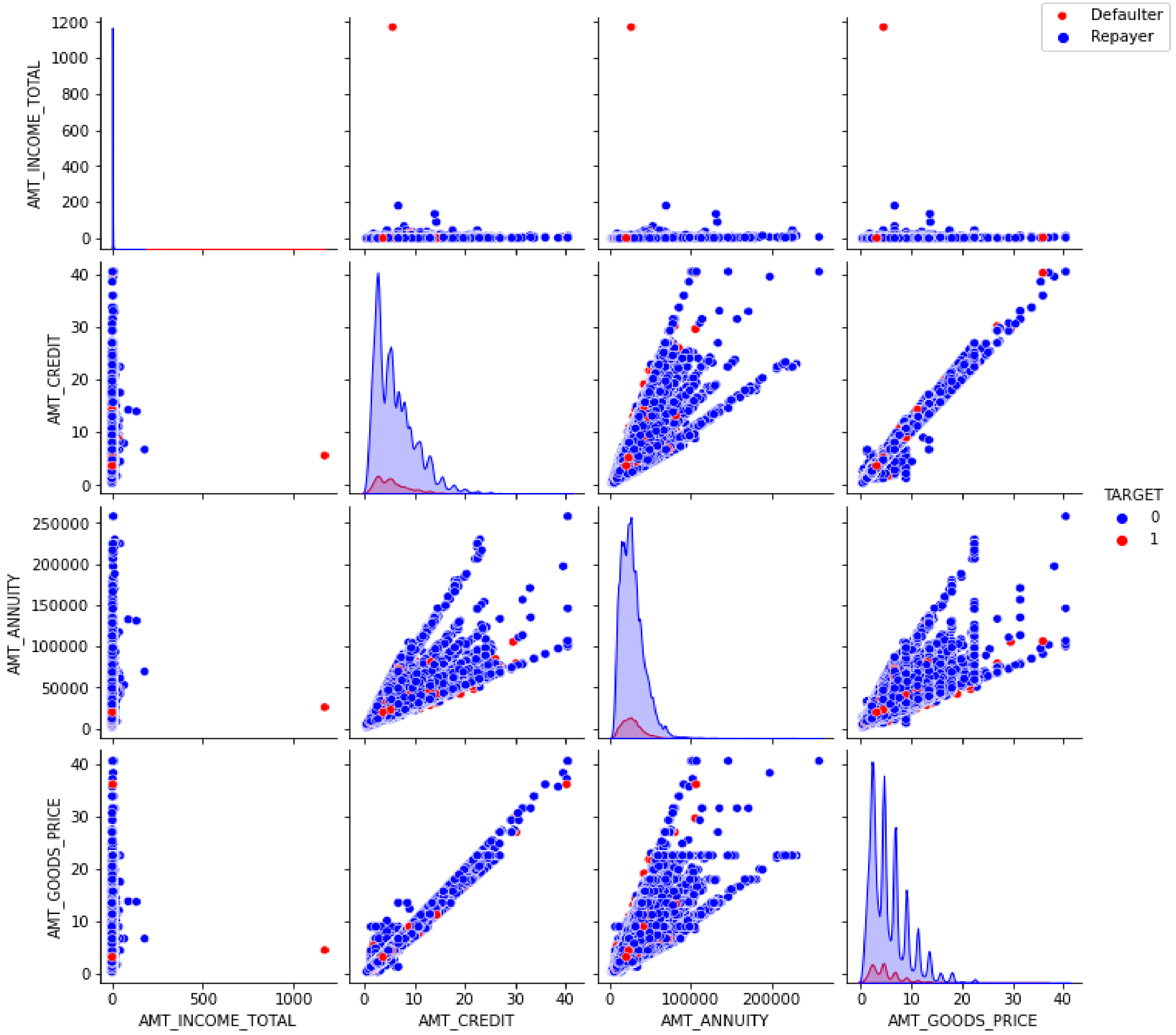


Inferences:

1. Loans are mostly given for goods price which are below 10 lakhs.
2. Most people pay annuity below 50K for the credit loan.
3. Credit amount of the loan is mostly less than 10 lakhs.
4. The re-payers and defaulters' distribution overlap in all the plots and hence we cannot use any of these variables on their own to make a decision.

1.2 **Numerical Bivariate Analysis:**

Plotting pair-plot between amount variable to draw reference against loan repayment status.



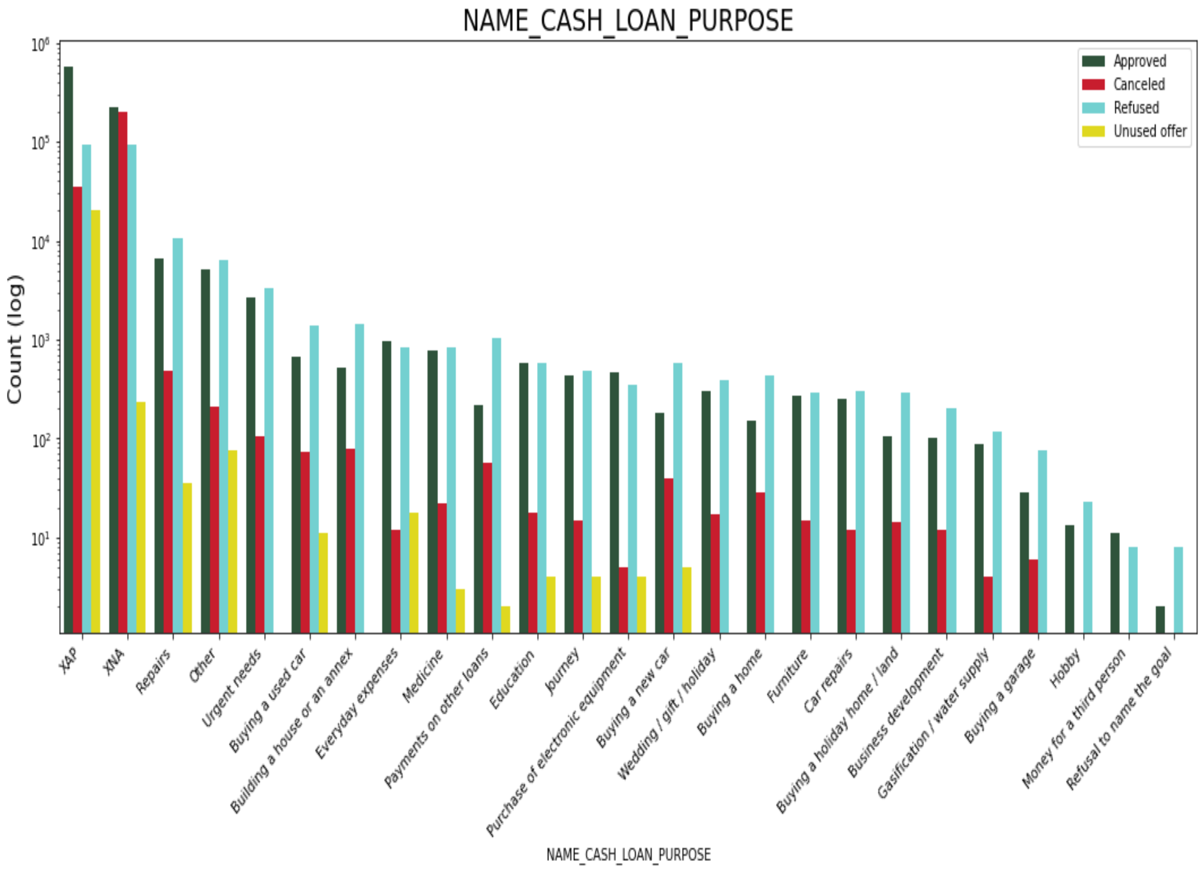
Inferences:

1. When Annuity Amount > 15K and Goods Price Amount > 20 Lakhs, there is a lesser chance of defaulters.
2. Loan Amount (AMT_CREDIT) and Goods price (AMT_GOODS_PRICE) are highly correlated as based on the scatterplot where most of the data are consolidated in form of a line.
3. There are very less defaulters for AMT_CREDIT >20 Lakhs

MERGED DATAFRAMES ANALYSIS

UNIVARIATE AND BIVARIATE ANALYSIS:

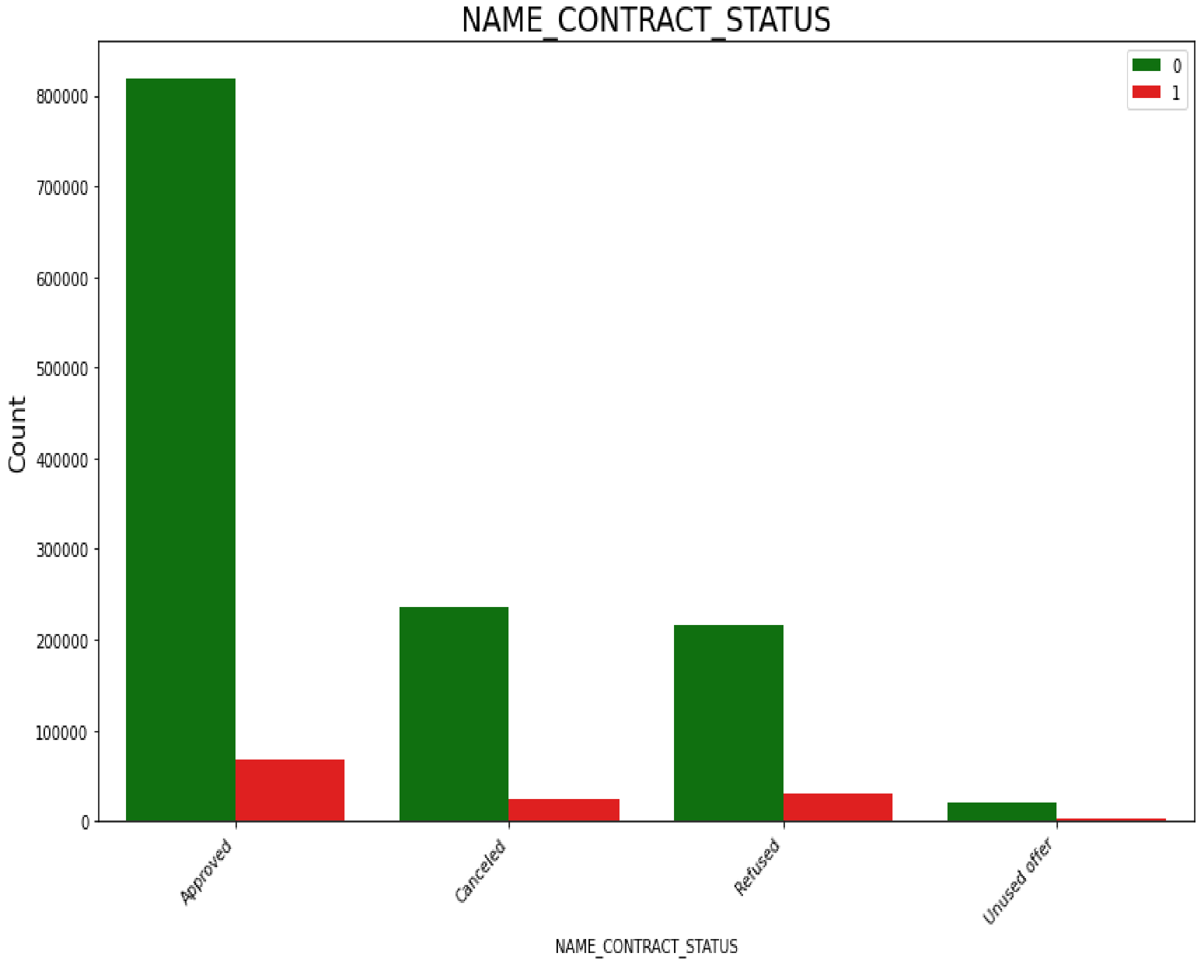
Plotting Contract Status vs purpose of the loan for re-payers.



Inferences:

1. Loan purpose has high number of unknown values (XAP, XNA).
2. Loan taken for the purpose of Repairs looks to have highest default rate.
3. Huge number application has been rejected by bank or refused by client which are applied for Repair or Other. From this we can infer that repair is considered high risk by bank.

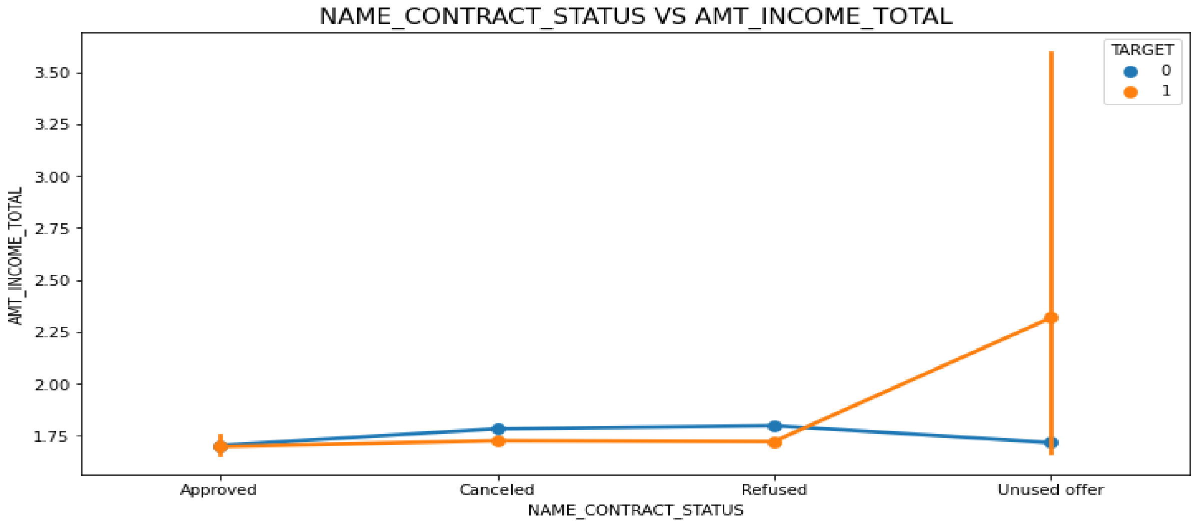
Checking Contract Status based on loan repayment status whether there is any business loss or financial loss.



Inferences:

1. 90% of the previously cancelled client have actually re-paid the loan. Revising the interest rates would increase business opportunity for these clients.
2. 88% of the clients who have been previously refused a loan has paid back the loan in current case.
3. Refusal reason should be recorded for further analysis as these clients could turn into potential repaying customer.

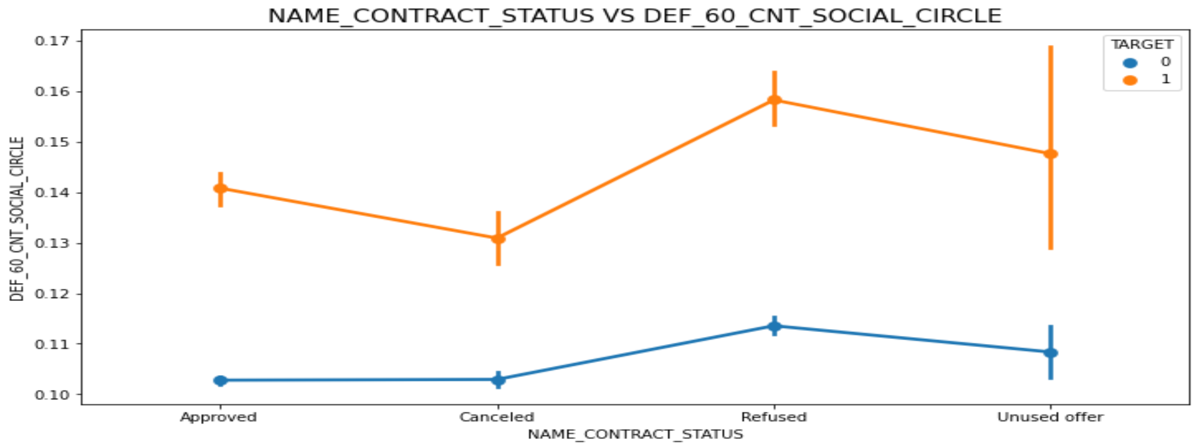
Plotting the relationship between income total and contact status.



Inferences:

The plot shows that the people who have not used offer earlier have defaulted even when their average income is higher than others.

Plotting the relationship between people who defaulted in last 60 days being in client's social circle and contact status.



Inferences:

Clients who have average of 0.13 or higher tend to default more and thus analysing client's social circle could help in disbursement of the loan.

CONCLUSION & RECCOMENDATIONS

Factor whether an applicant will be Re-payer:

- Academic degree has less defaults.
- Student and Businessmen have no defaults.
- RATING 1 is safer.
- Clients with Trade Type 4 and 5 and Industry type 8 have defaulted less than 3%
- People above age of 50 have low probability of defaulting
- Clients with 40+ year experience having less than 1% default rate
- Applicant with Income more than 700,000 are less likely to default
- Loans bought for Hobby, buying garage are being re-paid mostly.
- People with zero to two children tend to repay the loans.

Factor whether an applicant will be Defaulter:

- Men are at relatively higher default rate
- People who have civil marriage or who are single default a lot.
- People with Lower Secondary & Secondary education
- Clients who are either at Maternity leave OR Unemployed default a lot.
- People who live in Rating 3 has highest defaults.
- Avoid Low-skill Laborers, Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff as the default rate is huge.
- Organizations with highest percent of loans not repaid are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%). Self-employed people have relative high defaulting rate, and thus should be avoided to

be approved for loan or provide loan with higher interest rate to mitigate the risk of defaulting.

- Avoid young people who are in age group of 20-40 as they have higher probability of defaulting
- People who have less than 5 years of employment have high default rate.
- Client who has children equal to or more than 9 default 100% and hence their applications are to be rejected.
- When the credit amount goes beyond 3M, there is an increase in defaulters.

Recommendations:

- When we see the whether the previously rejected applicant has repaid the loan or not then we see a positive response of 90%. Hence recommending to note down the reason of previous rejection and provide the loan by slightly increasing the interest rate.
- Applicant who are about to achieve their superannuation age at work place has less default ratio. Therefore, providing them loan would be safer to increase the business.
- Scrutinize the loan application of those who live in rented house or with their parents because high ratio of loan defaulters is from this group.
- People with loan amount of 300-600k are high loan defaulter, hence increase the interest rate while providing the loan.
- Applicants with salary 3 lakhs and lesser have almost 90% default rate. Therefore, increase the interest rate while providing the loan.